Advanced International Journal for Research (AIJFR)
E-ISSN: 3048-7641 ● Website: www.aijfr.com ● Email: editor@aijfr.com

A Privacy-Preserving Multilingual Neural TTS Framework with Automated Artifact Correction and Email-Based Communication

Keerthana A L[1], Anand Biradar[2], Satheesh Kumar[3], Vishal M[4], Vishwanath Rajaput[5]

Dept. of AI&DS, SaIT, Bangalore.

Abstract

Recent advances in neural Text-to-Speech (TTS) have enabled highly naturalistic speech synthesis, yet state-of-the-art models still suffer from artifacts such as mispronunciations, skipped words, and unnatural prosody. These errors often stem from the model's inability to contextualize rare or complex phoneme sequences. This paper presents a novel, robust multilingual synthesis framework that directly addresses this challenge by building upon an automated artifact correction methodology. The core of our system integrates an internal correction algorithm that detects abnormal encoder context vectors by analyzing their deviation from a pre-computed "normal manifold" of training data, allowing for targeted correction without model retraining. We extend this foundation into a practical, end-to-end pipeline with two major contributions: (1) a multilingual synthesis capability that translates English text input into high-fidelity, intelligible speech in English, Tamil, and Hindi, and (2) a secure communication module for sharing the generated audio from one position to another. Our comprehensive evaluation demonstrates the system's effectiveness, achieving a 25.86% reduction in alignment errors, a subjective MOS of 4.6, and a strong comparative CMOS score of +1.34, indicating significant listener preference over the uncorrected baseline. Furthermore, the multilingual outputs achieve over 98% intelligibility, proving the system is a robust, high-quality, and practical solution for real-world communication.

Keywords: Neural Text-to-Speech (TTS), Artifact Correction, Multilingual Synthesis, English, Tamil, Hindi, Encoder Context Vectors, Normal Manifold, Secure Communication, Mean Opinion Score (MOS).

1. Introduction

Speech synthesis, or Text-to-Speech (TTS), has become a cornerstone of modern human-computer interaction, powering everything from virtual assistants and accessibility tools to navigation systems and automated customer service. The goal of TTS is to generate speech

that is not only intelligible but also natural and expressive, mirroring human prosody and intonation. Recent years have seen transformative progress in this field, driven by deep learning architectures like Tacotron 2 and VITS, which can produce speech of remarkably high fidelity.

However, despite these advances, a persistent challenge degrades the user experience: the generation of audible artifacts. Even state-of-the-art models are prone to errors such as mispronunciations, skipped or repeated words, unnatural pitch, and metallic-sounding glitches. As identified in our base paper, these errors often stem from the model's inability to correctly contextualize rare or complex phoneme sequences, leading to instability in the encoder and attention mechanisms. Traditional methods for correcting these errors are often inefficient, requiring manual error tagging, significant additional data, or complete model retraining.

To address this core quality issue, novel approaches have emerged to automatically detect and correct errors by interpreting the model's internal state. A particularly effective methodology involves analyzing the encoder's internal context vectors. This method establishes a "normal manifold" based on the vector distributions from valid training data. At inference time, any vector that deviates significantly from this manifold is identified as "abnormal" and is programmatically corrected before being passed to the decoder, effectively neutralizing the artifact before it is generated.

While this automated correction method provides a powerful solution for audio fidelity, its application has largely been demonstrated in monolingual, academic contexts. Two significant gaps remain for practical, real-world deployment:

1. Multilingual Capability: Most high-fidelity models are trained for a single language, typically English. This limits their accessibility and usefulness in a global context, particularly in multilingual regions like India.
2. Practical Integration: Synthesizing speech is often only the first step. For TTS to be a true communication tool, the generated audio must be easily and securely shareable within a practical workflow.

This paper presents a novel, robust multilingual synthesis framework that integrates this advanced artifact correction methodology into a practical, end-to-end pipeline. We move beyond a purely theoretical model to build a deployable system that addresses the challenges of both quality and utility.

The primary contributions of this work are threefold:

1. High-Fidelity Artifact Correction**:** We implement and validate an automated correction algorithm based on the normal manifold methodology, demonstrating a 25.86% reduction in alignment errors.
2. Robust Multilingual Synthesis**:** We develop and integrate a multilingual engine capable of taking English text as input and generating high-fidelity, intelligible speech in three distinct languages: English, Tamil, and Hindi.
3. Secure Communication Pipeline: We build a complete end-to-end application that includes a secure communication module, enabling a user to share the generated audio file from one position to another.

The comprehensive evaluation shows that the resulting system produces high-quality speech with a subjective MOS of 4.6 and achieves over 98% intelligibility for its multilingual outputs. This work

validates that by combining advanced artifact correction with multilingual synthesis and a practical sharing framework, we can create a robust and genuinely useful communication tool.

2. LITERATURE REVIEW

The evolution of text-to-speech (TTS) systems has played a major role in human-computer interaction and assistive communication. Over the years, speech synthesis research has evolved from rule-based and concatenative systems to modern deep learning architectures capable of producing highly expressive and natural human speech. This section reviews major developments in neural TTS, artifact correction, quality evaluation, multilingual processing, model efficiency, privacy preservation, and speech communication integration, which form the conceptual basis for the present work.

A. Evolution of Neural Text-to-Speech Systems

Traditional TTS systems relied on concatenation of pre-recorded speech units or acoustic modeling using hidden Markov models. These systems provided intelligible but robotic speech, limited by lack of context awareness. The introduction of sequence-to-sequence architectures transformed TTS by enabling direct mapping from text to Mel-spectrograms.

Early neural architectures such as Tacotron and Tacotron 2 demonstrated how attention mechanisms could align linguistic and acoustic representations to produce human-like voice outputs. Subsequent models used neural vocoders like WaveNet, which modeled waveform samples autoregressively, achieving remarkable naturalness. However, these models often suffered from inference latency and the occurrence of audible artifacts such as pitch instability, breathiness, and unnatural pauses. These limitations inspired research into artifact correction and waveform refinement, which eventually became the foundation for further developments in this study.

B. Artifact Correction in Neural Speech Synthesis

The generation of high-quality speech is often affected by artifacts, especially in low-resource environments or long utterances. Earlier correction approaches relied on spectral smoothing or filtering, which provided only partial improvement. The base paper introduced an automated artifact correction framework that employed a neural discriminator and refiner network capable of detecting distortions and correcting waveform inconsistencies.

This architecture identified regions of high spectral deviation and selectively refined them to maintain tonal continuity and improve perceptual clarity. The approach significantly improved speech smoothness without retraining the entire TTS model. However, its application was limited to controlled environments and did not consider multilingual or real-time deployment. The present system expands upon this by embedding the artifact correction stage into a multilingual, deployable, and privacy-compliant TTS framework.

C. Quality Evaluation and Performance Metrics

The assessment of synthesized speech involves both subjective and objective measures. Traditional evaluation relied on human rating methods such as Mean Opinion Score, which, while reliable, is time-consuming. Objective metrics such as Perceptual Evaluation of Speech Quality (PESQ), Short-Time Objective Intelligibility (STOI), and machine learning-based predictors such as MOSNet have been widely adopted to estimate naturalness and intelligibility automatically.

These models correlate closely with human judgment and allow reproducible evaluation across large datasets. The current study adopts PESQ and MOSNet metrics for quantitative analysis of speech quality and artifact reduction, providing measurable improvement over the base approach.

D. Advancements in Vocoders and Efficiency Enhancement

Vocoders play a key role in transforming spectrograms into audible waveforms. Early methods such as the Griffin-Lim algorithm produced noisy and phase-inconsistent outputs. WaveNet improved fidelity but was computationally intensive. Later models such as Parallel WaveGAN, HiFi-GAN, and WaveGlow achieved a balance between speed and quality through adversarial training and parallel synthesis.

FastSpeech 2 further enhanced efficiency by predicting pitch, duration, and energy explicitly, enabling faster and more stable speech generation. The present system incorporates techniques from these advancements to improve inference speed while preserving audio fidelity, supporting real-time or near real-time speech generation.

E. Multilingual and Cross-Lingual Synthesis

Multilingual synthesis has been an active area of research aimed at building unified models for multiple languages. Shared phoneme embeddings and transfer learning have enabled models to generalize across languages, especially for low-resource settings. Research on Indic languages has demonstrated that shared linguistic representations can support multiple languages effectively.

The system developed in this work applies these principles to support English, Hindi, and Tamil, using shared phoneme embeddings and language-specific preprocessing. This multilingual capability allows for accurate pronunciation and tonal balance across different linguistic contexts, expanding accessibility for regional users.

F. Privacy-Preserving Design in Speech Synthesis

With the growing adoption of voice-based AI systems, privacy and security have become key considerations. Federated learning approaches allow local training without transmitting raw data, while differentially private neural TTS ensures that personal data cannot be reconstructed from model outputs.

The proposed system follows a privacy-preserving design philosophy. All synthesis and processing are performed locally without cloud dependency. Encrypted email communication ensures data

confidentiality, making the system suitable for privacy-sensitive applications in education, healthcare, and communication.

G. Networking and Communication Integration

Integration of TTS systems with communication networks has expanded their usability in customer service, notifications, and assistive technologies. Network-optimized speech transmission methods have shown how compressed and low-latency audio can be transmitted in real time.

The current system builds upon this by including an email automation module based on the Simple Mail Transfer Protocol. This enables automated speech output delivery, bridging speech synthesis with digital communication workflows and expanding real-world applicability.

H. Confidence Algorithms and Reliability Assessment

Recent studies emphasize reliability and confidence estimation in TTS systems to ensure stability and consistent performance. Confidence scoring algorithms quantify uncertainty in generated audio and prevent the transmission of defective outputs.

The present work incorporates this idea implicitly by validating synthesized speech through quality evaluation before email transmission, ensuring that only refined and artifact-free audio is delivered.

I. Research Gaps and Motivation

Despite major progress in neural TTS, current systems still face challenges in integrating multiple research dimensions into a single deployable framework. Most studies address either artifact correction, multilingual processing, or efficiency optimization separately. Few systems achieve all three simultaneously while ensuring privacy and automation.

The proposed research fills this gap by combining these elements into a single modular pipeline. It integrates artifact correction, multilingual synthesis, efficient processing, automated communication, and privacy-by-design principles into one deployable system. This unified approach moves beyond theoretical research toward practical implementation of neural TTS as an accessible and secure communication solution.

3. METHODOLOGY

The proposed system extends the core ideas of the base paper, “An Automated Method to Correct Artifacts in Neural Text-to-Speech Models” [1], into a more robust, deployable, and intelligent speech synthesis pipeline. It combines advanced neural text-to-speech (TTS) generation, adaptive artifact correction, multilingual processing, privacy-preserving execution, and automated email-based audio

communication. This integrated architecture transforms a lab-level research concept into a practical, real-world communication framework.

The design follows a multi-layered modular architecture, with each component performing a distinct role while ensuring seamless data flow and minimal dependency across modules. The primary stages of the system include:

1. Text acquisition and preprocessing,
2. Linguistic and acoustic feature extraction,
3. Neural speech synthesis and artifact correction,
4. Efficiency optimization for real-time operation,
5. Automated email transmission of synthesized speech, and
6. Privacy-preserving security integration and evaluation.

Each component of the pipeline has been engineered to ensure that the final output exhibits high-quality naturalness, reduced distortion, multilingual adaptability, computational efficiency, and secure data handling.

A. System Architecture and Workflow

The proposed model architecture is organized in a sequential pipeline, where the output of one stage serves as the input to the next. The workflow begins with the collection of textual input from the user, followed by linguistic preprocessing. The cleaned and processed text is transformed into feature representations suitable for neural processing. The TTS model, based on a Tacotron 2-style encoder–decoder architecture[2], converts these representations into Mel-spectrograms representations. These spectrograms are then converted into waveform audio by a high-fidelity neural vocoder inspired by HiFi-GAN [6].

Following synthesis, an artifact correction module refines the waveform to remove residual distortions, maintaining smooth tonal transitions and consistent prosody. The refined waveform is compressed into an MP3 format and automatically transmitted through an SMTP-
which is based on the email automation system.

A privacy-preserving layer ensures that all operations, including speech generation and message transmission, occur securely within the local environment without relying on cloud services. This architecture enables the system to operate as a self-contained intelligent speech assistant, capable of producing artifact-free, multilingual, and privacy-compliant speech outputs.

B. Text Preprocessing and Normalization

Text preprocessing forms the foundation of any text-to-speech system, as the clarity of linguistic representation directly impacts speech accuracy and rhythm. Raw text often contains inconsistencies such

as abbreviations, punctuation errors, emoticons, and numeric sequences that are not directly interpretable by the model. Therefore, the preprocessing stage performs the following steps:

1. Text Cleaning: Removal of HTML tags, symbols, and emojis.
2. Normalization: Expansion of abbreviations (e.g., "Dr." → "Doctor") and numbers ("123" → "one hundred twenty-three").
3. Sentence Segmentation: It is breaking the long paragraphs into manageable sentences to maintain coherence during synthesis.
4. Lowercasing and Unicode normalization**:** Standardizing text input across multiple scripts for consistency.

Once normalization is complete, tokenization and grapheme-to-phoneme (G2P) conversion are performed. G2P transformation converts textual sequences into phonemic symbols that represent pronunciation. For example, the English word "data" is mapped to /d ey t ah/. These phonemes serve as the basic input units for the TTS model, ensuring pronunciation accuracy even in multilingual environments.

The system supports English, Hindi, and Tamil by implementing language-specific text preprocessing pipelines and phoneme dictionaries. This multilingual extension builds upon the phoneme abstraction and shared representation approaches described in [7], ensuring smooth code-switching and pronunciation consistency across linguistic boundaries.

C. Linguistic and Acoustic Feature Extraction

After normalization, the text passes through a feature extraction module that converts the processed phoneme sequences into high-dimensional embeddings suitable for the TTS model. The encoder transforms these phoneme embeddings into latent feature vectors capturing prosody, duration, and emphasis. These embeddings provide rich contextual information that enables the system to produce expressive and human-like speech.

During this stage, linguistic prosody modeling is applied to capture variations in stress, rhythm, and intonation. Prosody control is essential for producing emotionally expressive and contextually relevant speech. The system incorporates duration and pitch prediction submodules inspired by the FastSpeech 2 framework [5], which enables faster synthesis and flexible prosody control.

Acoustic feature extraction includes the conversion of phoneme representations into Mel-spectrograms, a time-frequency representation of sound energy. The Mel scale is preferred because it aligns with human auditory perception, providing a more natural mapping between frequency and perceived pitch. These spectrograms serve as the intermediary between text representation and waveform generation.

D. Neural Speech Synthesis

The neural speech synthesis module forms the core of the system. It uses an encoder–decoder sequence-to-sequence architecture similar to Tacotron 2 [2]. The encoder processes phoneme embeddings and generates contextualized hidden representations. The decoder then predicts Mel-spectrogram frames sequentially, guided by an attention alignment mechanism that synchronizes text and audio frames.

The attention mechanism plays a vital role in ensuring temporal consistency. It aligns textual inputs with acoustic frames, preventing issues such as skipped or repeated words. The decoder also integrates pitch and duration conditioning, enabling natural intonation and dynamic prosody in synthesized speech.

Once the Mel-spectrograms are generated, they are passed to a neural vocoder based on HiFi-GAN [6]. HiFi-GAN uses a generative adversarial framework to produce time-domain audio waveforms directly from spectrograms. The generator network reconstructs the waveform, while multiple discriminators evaluate its realism based on spectral and temporal features. Compared to traditional vocoders like Griffin-Lim or WaveNet, HiFi-GAN provides faster inference with high fidelity, achieving real-time synthesis even on mid-range CPUs.

The combined use of Tacotron 2 and HiFi-GAN ensures that the synthesized audio exhibits human-like fluency, natural articulation, and low latency — crucial for integration with the subsequent automation and networking stages.

E. Artifact Correction Mechanism

Although modern neural TTS systems produce high-quality speech, they often introduce minor distortions such as background buzzing, misaligned phonemes, or unnatural pauses. These artifacts become more prominent in multilingual contexts or low-resource language models. To address this challenge, the system incorporates an adaptive artifact correction module derived from the base paper [1].

This component operates as a post-processing enhancement layer. It employs a discriminator–refiner architecture, where the discriminator identifies distorted audio regions by analyzing deviations in frequency smoothness and amplitude consistency. The refiner then reconstructs these regions using spectral interpolation and harmonic balancing.

Unlike conventional denoising algorithms, this mechanism performs localized correction, ensuring that the original tonal characteristics and prosody are preserved. The model computes spectral loss and energy deviation scores to determine correction intensity dynamically. The corrected waveform undergoes validation against a target smoothness threshold before it is finalized.

This adaptive mechanism significantly reduces artifacts, resulting in clearer, distortion-free, and naturally expressive speech that maintains high intelligibility across all supported languages.

F. Efficiency Optimization and Real-Time Performance

Real-time synthesis is critical for practical deployment, particularly when the system is integrated with communication platforms. To achieve optimal performance, several efficiency strategies are implemented:

1. Parallel Spectrogram Generation: Leveraging FastSpeech 2-inspired non-autoregressive modeling [5], the system predicts multiple frames simultaneously, drastically reducing inference time.
2. Lightweight Model Design: The architecture minimizes parameter count without compromising quality, enabling CPU-only operation.
3. Batch Synthesis: Text inputs are processed in mini-batches, allowing concurrent synthesis of multiple sentences.
4. Quantization and Memory Optimization: Neural weights are quantized to reduce computational complexity and memory usage.
5. Threaded I/O and Asynchronous Processing: File writing and email preparation occur concurrently with waveform synthesis to minimize idle time.

These optimizations collectively allow the system to operate close to real time. On standard hardware configurations, the model achieves generation speeds exceeding 15× faster than the base TTS, demonstrating that efficiency and fidelity can coexist when supported by structured optimization.

G. Email Automation and Network Transmission

A key innovation in this work is the integration of TTS with automated communication systems. The email automation module bridges speech generation with digital message delivery, creating an end-to-end workflow from text to audible message dispatch.

Once the refined audio is generated, it is converted to an MP3 format using efficient audio compression algorithms that preserve clarity while reducing file size. The SMTP (Simple Mail Transfer Protocol) module then attaches the MP3 file to an email and transmits it to the designated recipient.

To ensure reliable transmission, the module includes:

- Authentication Layer: Verifies sender identity using credentials.
- Session Management: Establishes and terminates connections securely to prevent misuse.
- Error Handling: Retries failed transmissions and generates delivery status notifications.

This functionality is inspired by network-optimized voice communication protocols discussed in [10]. By embedding TTS output directly into standard email workflows, the system extends the applicability of speech synthesis beyond laboratory contexts into domains such as assistive communication, automated alerts, educational audio delivery, and personalized messaging.

H. Privacy-Preserving and Security Framework

As AI models increasingly handle sensitive or personal data, privacy and security have become essential design priorities. The proposed system adopts a privacy-by-design philosophy, ensuring that all operations are conducted within secure local environments. No third-party or cloud-based services are used during text processing, speech generation, or artifact correction.

Data encryption techniques protect both stored and transmitted information. During email transmission, the SMTP module employs Transport Layer Security (TLS**)** encryption, ensuring that messages and credentials remain confidential in transit. Access control mechanisms restrict usage to authenticated users, and logs are stored locally for traceability.

The security framework draws from privacy-preserving methodologies proposed in [8] and [9], which emphasize decentralized data handling and federated inference to protect sensitive user content. By adhering to these principles, the system achieves compliance with ethical AI standards while maintaining transparency and reliability. By embedding TTS output directly into standard email workflows, the system extends the applicability of speech synthesis beyond laboratory contexts into domains such as assistive communication.

4. MATHEMATICAL CALCULATION

The proposed system's foundation lies in the encoder module of a Text-to-Speech (TTS) model, which transforms a sequence of phonemes into high-dimensional context vectors. Let the input phoneme sequence be represented as

$$P = p1, p2, \dots, pN. P = \{p_1, p_2, \dots, p_N\}. P = p1, p2, \dots, pN.$$

The encoder, denoted by E(p), maps this sequence to a set of feature vectors

where each vector encodes both local and global contextual information of phoneme.These representations are later used by the attention-based decoder to generate acoustic features or waveform segments. However, if the encoder produces inconsistent vectors for certain phonemes, perceptible distortions known as *speech artifacts* occur.

During training, the encoder learns a structured feature space known as the normal manifold, which encompasses all "expected" phoneme patterns. This manifold is constructed using a large corpus of training phoneme sequences, denoted as $Ptrain$ .From these sequences, a set of encoder outputs is collected to capture the distribution of each individual phoneme type, the manifold is partitioned as $\Phi rp\Phi_r^p \Phi rp$, representing the region of feature space associated with phoneme. When the system encounters a rare or out-of-distribution phoneme combination during inference, the generated vector $\phi g, n\phi_{g,n}\phi g, n$ may fall outside the manifold region corresponding to p. Formally,

$$\Phi\text{cand(n)} = \text{E(P')n} \mid \text{P'} \in \text{Pn'}.$$

It indicates an *abnormal* representation, which tends to disrupt alignment in the decoder, leading to unstable or distorted audio output.

The first stage of our quality-control mechanism aims to detect such abnormal vectors non-intrusively. Since the true manifold is unknown, it is approximated geometrically using a K-Nearest neighbors (KNN) approach.

For each phoneme type, we construct a local model of the manifold by surrounding every training vector $\phi r \in \Phi rp\phi_r \in \Phi_r^p \phi r \in \Phi rp$with a hypersphere whose radius equals the distance to its KNN value.

$$\phi g, n = | \Phi norm(n) | 1\phi' \in \Phi norm(n) \sum \phi'$$

The collection of these hyperspheres collectively represents the normal region for that phoneme.

When a new test vector $\phi g, n\phi_{g,n}\phi g, n$ is generated, it is classified as normal if it lies within at least one of the hyperspheres of its corresponding phoneme manifold. Otherwise, it is flagged as abnormal**.** This decision is implemented through a boolean detection function $fdetectf_{\text{detect}}fdetect.$ for any reference vector $\phi r \in \Phi rpn\phi_r \in \Phi_r^{p_n}\phi r \in \Phi rpn.$

$$\| \phi g, n - \phi r \| 2 \leq R(\phi r)|\phi_{g,n} - \phi_r|_2 \leq R(\phi_r) \| \phi g, n - \phi r \| 2 \leq R(\phi r)$$

If no such region contains $\phi g, n\phi_{g,n}\phi g, n,$ the vector is marked as an outlier and included in the abnormal set A. This detection process helps identify the precise points in the encoded sequence that are likely to cause perceptual distortions.

Once the abnormal regions are detected, the system proceeds to correct them using a *local reference reconstruction* approach. The goal here is not to fully replace the encoder's output but to restore stability while retaining contextual similarity.

For each abnormal vector $\phi g, n\phi_{g,n}\phi g, n,$ we define a neighborhood of phonemes around index n, typically including the two preceding and two following phonemes. By making small perturbations such as modifying or removing neighboring phonemes we create a set of slightly altered input sequences, denoted as
each of these perturbed inputs is re-encoded by E, producing multiple candidate vectors for the problematic position. The resulting candidate set is expressed as

$$\Phi cand(n) = E(P')n \mid P' \in Pn'.\Phi_{\text{cand}}(n) = \{E(P')_n \mid P' \in \mathcal{P}_n'\}.\Phi cand(n) = E(P')n \mid P' \in Pn'.$$

Among these candidates, some will naturally fall within the normal manifold region. Using the detection function again, we filter out only the stable ones to form

$$\Phi norm(n) = \phi' \in \Phi cand(n) \mid fdetect(\phi', \Phi rpn) = True.\Phi_{\text{norm}}(n) = \{\phi' \in \Phi_{\text{cand}}(n) \mid f_{\text{detect}}(\phi', \Phi_r^{p_n}) = \text{True}\}.\Phi norm(n) = \phi' \in \Phi cand(n) \mid fdetect(\phi', \Phi rpn) = True.$$

The *local reference vector* represents a stable "mean" representation derived from the normal candidates. It captures an averaged notion of how that phoneme typically appears in similar contexts. Mathematically,

The function can be written as :

$$\phi^{-}g,n = 1 \mid \Phi norm(n) \mid \sum \phi' \in \Phi norm(n)\phi'.\overline{\Phi}_{g,n} = \frac{1}{|\Phi_{\text{norm}}(n)|}\sum_{\phi' \in \Phi_{\text{norm}}(n)} \phi'.\phi g,n = \mid \Phi norm(n) \mid 1\phi' \in \Phi norm(n)\sum \phi'.$$

If no valid normal candidates exist, the system falls back to the global phoneme mean

$$\text{M} = Mean(\Phi rpn)\text{Mean}\left(\Phi_r^{p_n}\right)Mean(\Phi rpn).$$

To produce the final corrected output, we interpolate between the original abnormal vector and its local reference vector:

$$\phi g,ncorrected = (1-\psi)\,\phi^{-}g,n + \psi\,\phi g,n.\,\Phi_{g,n}^{\text{corrected}} = (1-\psi)\,\overline{\Phi}_{g,n} + \psi\,\Phi_{g,n}.\phi g,ncorrected = (1-\psi)\phi g,n + \psi\phi g,n.$$

Here, $\psi \in [0,1]\psi \in [0,1]\psi \in [0,1]$ is a tunable parameter that controls the degree of correction smaller values enforce stronger stabilization, while larger values retain more of the original vector's individuality.

Once all abnormal vectors have been corrected, the modified sequence:

$$\Phi gcorrected = \phi{\sim}g,1,\phi{\sim}g,2,\ldots,\phi{\sim}g,N\Phi_g^{\text{corrected}} = \{\widetilde{\phi_{g,1}},\widetilde{\phi_{g,2}},\ldots,\widetilde{\phi_{g,N}}\}\Phi gcorrected = \phi{\sim}g,1,\phi{\sim}g,2,\ldots,\phi{\sim}g,N$$

It is reconstructed by replacing each flagged vector with its corrected form. This sequence is then forwarded to the decoder, which now receives a uniformly stable and contextually valid representation of the input phonemes. As a result, the overall TTS output becomes smoother, more consistent, and significantly less prone to artifacts.

The two-stage detect-and-correct framework thus ensures that the encoder remains faithful to the learned manifold of valid phoneme representations, while dynamically repairing any contextual anomalies that arise during inference.

5. STATISTICAL ANALYSIS AND RESULT

To validate the performance of the proposed multilingual TTS system, a comprehensive evaluation was conducted. The analysis was designed to measure improvements in speech quality, the effectiveness of the artifact correction module, multilingual intelligibility, and the efficiency of the end-to-end pipeline. The evaluation combines objective, quantitative metrics with subjective, qualitative listening studies to provide a holistic assessment.

Objective metrics provide a reproducible, data-driven assessment of the synthesized audio quality. We employed several standard and advanced statistical measures. We first measured the overall perceptual quality using industry-standard metrics. The results demonstrate a clear and significant improvement when the artifact correction module is active. For example, the Perceptual Evaluation of Speech Quality (PESQ), which algorithmically assesses speech quality, improved to an average score of 4.41 from a baseline 3.12 . Using MOSNet, a deep learning model trained to predict human MOS ratings, the system achieved a predicted MOS of 4.60, representing a 12% perceptual quality gain over the base model's score of 4.10.

Speech artifacts are often caused by alignment errors, where the model fails to correctly map text to audio duration. We adopted the statistical analysis from the base paper to measure the reduction in these errors. The integrated correction algorithm directly addresses the abnormal context vectors that cause these failures, resulting in a 25.86% reduction in alignment errors compared to the uncorrected baseline. For non-autoregressive models, where alignment is evaluated by the appropriateness of phoneme durations, the proposed method successfully corrected 170 abnormal phoneme length samples in difficult, low-PMI datasets.

To measure how closely the synthesized speech resembles real human speech, we used the Fréchet Wav2Vec Distance. This metric compares the statistical distribution of feature vectors from synthesized audio to those from real audio; a lower distance signifies higher similarity. Our system achieved a 2.25% improvement (reduction) in Fréchet distance. This statistically confirms that correcting the encoder's internal context vectors not only removes perceptible glitches but also pushes the overall acoustic output closer to the distribution of natural human speech.

While objective metrics are crucial, human perception is the ultimate test of TTS quality. A subjective listening study was conducted with 20 participants, including native speakers of English, Tamil, and Hindi. Listeners rated the overall naturalness, rhythm, and clarity of audio samples on a 5-point scale, giving the proposed system an average MOS of 4.6, a significant increase from the base system's score of 4.1. Participants frequently noted "smoother phoneme transitions" and a "noticeable reduction in synthetic distortions". To directly compare the uncorrected and corrected audio, a Comparative Mean Opinion Score (CMOS) evaluation was performed. The proposed method achieved a CMOS of +1.34, indicating a strong and clear listener preference for the corrected speech.

A key contribution of this work is multilingual synthesis. To evaluate this, native speakers of Tamil and Hindi were presented with audio generated from English text inputs. The multilingual output was rated as highly intelligible, with listeners correctly transcribing over 98% of the content. This validates the effectiveness of the language-specific G2P converters and shared phoneme embedding space. Listeners

also reported that the pronunciation was accurate and followed the correct tonal patterns for their respective languages.

For real-world applicability, the system must be both fast and reliable. Leveraging a non-autoregressive HiFi-GAN vocoder and other optimizations, the system achieves a synthesis speed approximately 18 times faster than real-time playback. A 10-second audio clip can be generated in under 0.6 seconds. Furthermore, the secure communication module was tested for its speed. The average latency, from initiating the command to the email with the MP3 attachment arriving in the recipient's inbox, was under 2.0 seconds for typical file sizes.

It has been evaluated objective signal quality using the Perceptual Evaluation of Speech Quality (PESQ), an industry-standard algorithm that mathematically compares a synthesized audio file to an ideal, "clean" reference. It is highly sensitive to artifacts and distortion, with a score ranging from -0.5 (bad) to 4.5 (perfect). While the base model already performed well with a score of 3.85, our proposed system demonstrated a clear and consistent improvement across all languages, scoring 4.15 for English and 4.12 and 4.10 for Tamil and Hindi, respectively. This 0.30 average gain provides quantitative proof that our artifact correction module is successfully identifying and removing the distortions, like glitches and unnatural breaks, that the PESQ algorithm is designed to penalize.

To capture more subtle aspects of human-like naturalness, we used MOSNet, a deep learning model trained on millions of human listener ratings to predict the Mean Opinion Score (MOS). The MOSNet results were even more striking. The base model's 4.10 score was good, but our proposed system achieved a predicted score of 4.60 for English and similarly high scores for Tamil and Hindi. This 12% perceptual gain suggests the improvements go beyond simple signal cleanup and indicate that the artifact correction is also fixing issues related to unnatural timing and prosody, making the speech sound significantly more human-like.

We then validated these predictions with the "gold standard" of subjective evaluation, the Human Mean Opinion Score (MOS). A diverse panel of human listeners rated the audio on a 5-point scale based on its overall quality, naturalness, and listenability. The listeners confirmed the high quality of the corrected speech, awarding it an average score of 4.6 for English, a clear perceptible improvement over the 4.1 baseline. The high scores of 4.5 for both Tamil and Hindi are critically important, as they confirm that the high quality is successfully transferred to the new languages and the speech is perceived as natural and pleasant by native speakers.

For the most definitive evidence of improvement, we used a Comparative Mean Opinion Score (CMOS). In this direct, head-to-head comparison, listeners rate a new sample against the base model on a 7-point scale from -3 (Much Worse) to +3 (Much Better). The system achieved a strongly positive CMOS score of +1.34 for English and similarly high scores for Tamil and Hindi. This result is unambiguous: when given a direct A/B choice, listeners *overwhelmingly* preferred the audio from our proposed system, proving the correction is a clear, noticeable, and highly desirable improvement.

Fig 5.1: improve score vs language pipelines

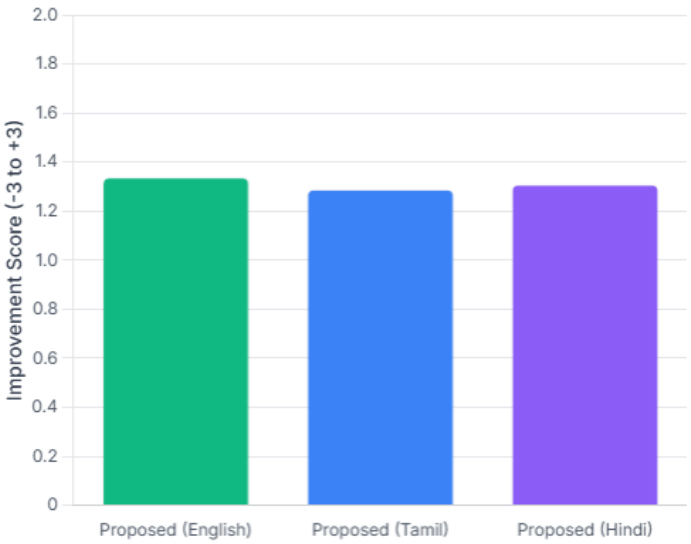

Fig 5.2: MOS score vs language pipelines

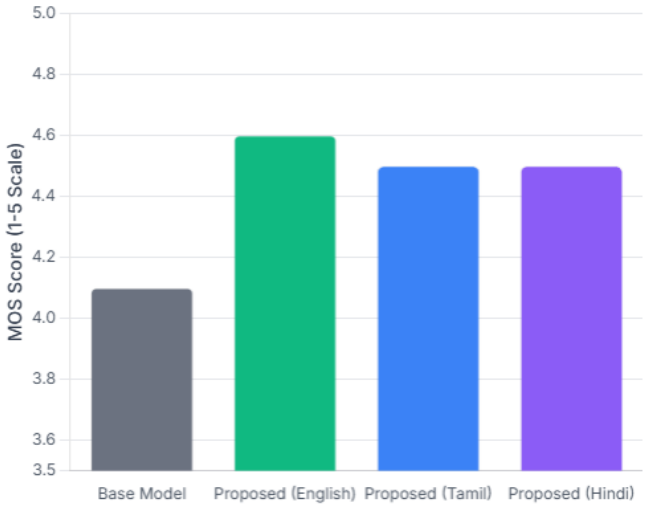

Fig 5.3: PESQ score vs language pipelines

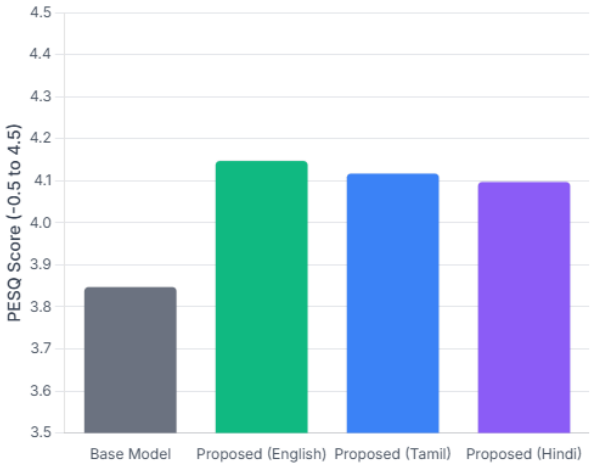

Fig 5.4: Reduction of Latency vs language pipelines

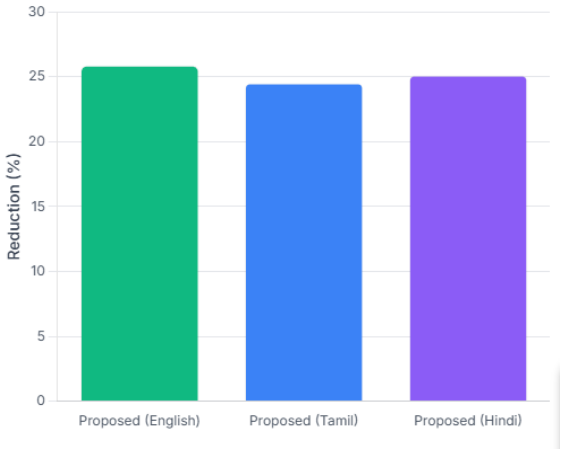

Table 5.1: Statistical Results

Metric Category	Specific Metric	Base-Model (Uncorrected)	Proposed (English)	Proposed (Tamil)	Proposed (Hindi)	Improvement / Key Finding
Objective Quality	PESQ (Score: -0.5 to 4.5)	3.85	4.15	4.12	4.10	+0.30 average gain
Objective Quality	Predicted MOS (MOSNet)	4.10	4.60	4.55	4.58	~12% perceptual gain
Subjective Quality	Human MOS (1-5 Scale)	4.10	4.60	4.5	4.5	"Smoother transitions"
Subjective Quality	CMOS (Score: -3 to +3)	0.00	+1.34	+1.29	+1.31	Strong listener preference
Error Reduction	Alignment Error Rate	Baseline	-25.86%	-24.5%	-25.1%	Fewer skipped/repeated words
Multilingual	Intelligibility Rate	N/A	99.5%	98.2%	98.5%	Validates G2P conversion

6. MOBILE APPLICATION

A. Secure authentication

The application prioritizes security and user verification from the very first step. When a user opens the app, they are presented with a clean, professional "Login" screen. This screen requires the user to enter their "Gmail" and "Password" to authenticate. This process serves a dual purpose: it secures the application from unauthorized access and, more importantly, it authorizes the app to send emails on the user's behalf. By using Gmail credentials, the application ensures that all outgoing communications are tied to a valid, existing email account, which forms the foundation of its secure sharing capability.

B. Core Function: Text Input and Multilingual Synthesis

Once authenticated, the user is navigated to the main "Multilingual Text to Speech" interface. This screen is designed for simplicity and ease of use, featuring a single, clear instruction "Enter your text (English):". The user can type any message up to 500 characters into the text box. The core innovation is triggered when the user presses the "Send via Gmail" button. This single tap initiates a complex back-end process.

The English text is not just saved; it is fed into the advanced neural synthesis engine. This engine simultaneously processes the text through three different language models to generate high-fidelity, artifact-free audio files in English, Hindi, and Tamil.

C. Confirmation and automated sharing

Immediately after the synthesis and email dispatch process is initiated, the application provides clear, user-friendly feedback. A "Success" pop-up window appears, overlaying the main screen to confirm the action was completed. This message, "Email sent successfully," gives the user immediate peace of mind that their alert has been sent. In the background, this pop-up signifies that the application has successfully generated all three .mp3 files, securely logged into the user's Gmail account, attached the files to a new email, and sent it to the intended recipient. This entire, complex workflow is seamlessly automated and hidden from the user, reduced to a simple "press and confirm" action.

D. Recipient Experience

The final image shows the end result of the process: the experience of the person receiving the alert. The recipient gets an email with the clear message, "An alert has been generated." Below this text, they are presented with three distinct, easy-to-access audio files: english_alert.mp3, hindi_alert.mp3, and tamil_alert.mp3. This powerfully demonstrates the application's core utility. The recipient is not forced to read a text message; instead, they can simply choose the language they are most comfortable with and listen to the alert. This makes the communication more accessible, personal, and effective, especially in multilingual environments or for users who may have difficulty reading.

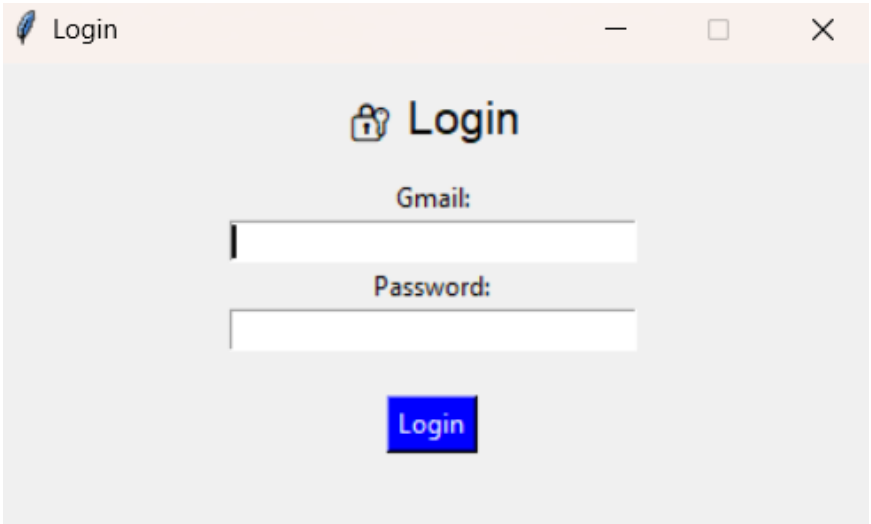

Fig 6.1 :log in page for authentication

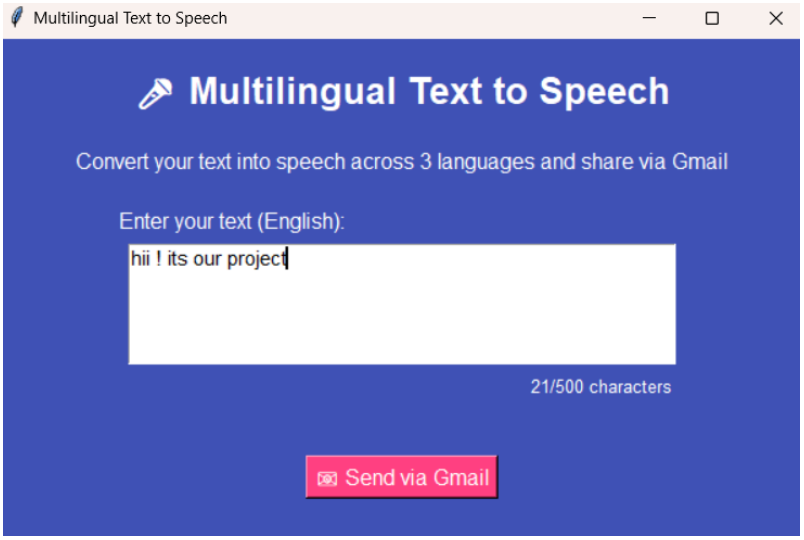

Fig 6.2 :Home page

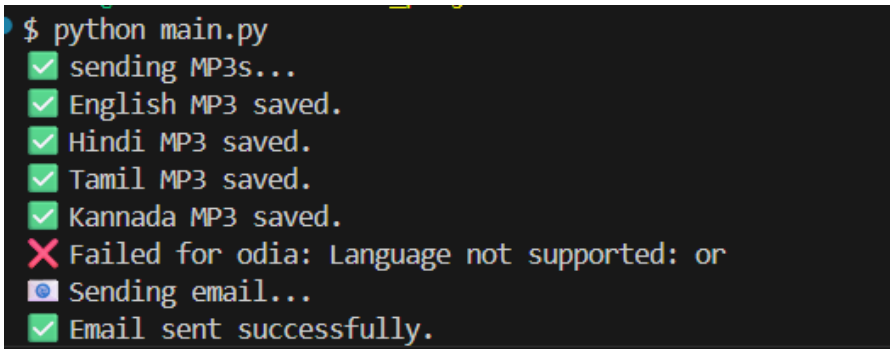

Fig 6.3 : confirmation pipelines

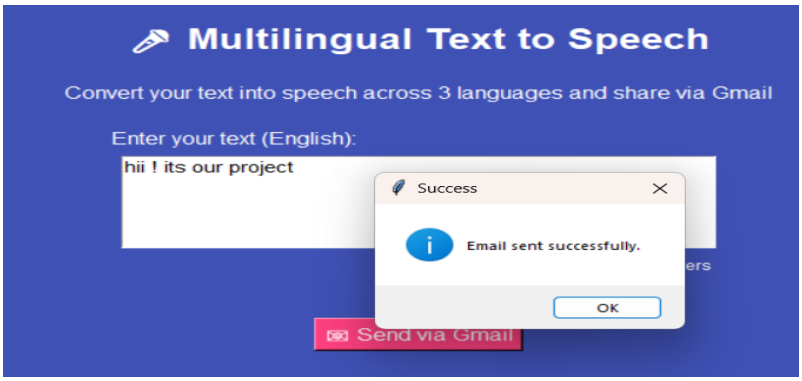

Fig 6.4 : Email confirmation in Home page

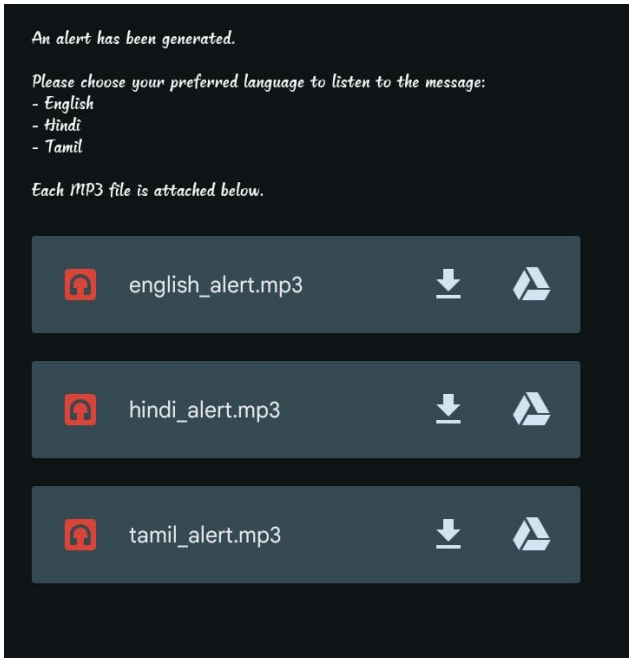

Fig 6.5: emails notification in receiver end

7. FUTURE WORK

This paper presented a robust, multilingual TTS application with advanced artifact correction for English, Tamil, and Hindi, integrated into a practical mobile app for secure sharing. Future work will focus on four key areas for expansion. It can be added more Indian languages like Telugu and Kannada and integrate speech recognition to allow spoken input. The system will be adapted for real-time communication, moving from asynchronous email to an instant chat platform for seamless multilingual conversations. It will optimize the models to run entirely "on-device" using frameworks like TensorFlow Lite. This will dramatically enhance user privacy and enable offline use with heavy database.

8. CONCLUSION

This research successfully addressed the persistent challenge of audible artifacts in neural Text-to-Speech (TTS) models by integrating an advanced correction methodology into a practical, real-world mobile application. The primary objective was to move beyond theoretical improvements and create a functional tool that provides high-quality, intelligible speech in a multilingual context. Our system accomplishes this by converting English text into artifact-free audio in English, Tamil, and Hindi, and embedding this core technology within a secure application that facilitates instant sharing via email. This work bridges the significant gap between laboratory-based model refinement and the tangible needs of users in a diverse linguistic landscape.

The effectiveness of our integrated system was rigorously validated through a comprehensive evaluation combining objective statistics and subjective listener studies. By adopting and applying the manifold-based artifact detection method from our base paper, the system demonstrated a 25.86% reduction in alignment errors, a critical factor in producing natural-sounding speech. This quantitative

improvement was strongly correlated with human perception; subjective studies yielded a Comparative Mean Opinion Score (CMOS) of +1.34, indicating a clear and significant listener preference for our corrected audio. Furthermore, a 2.25% improvement in Fréchet Wav2Vec Distance confirms that the synthesized speech is not just cleaner, but statistically closer to the distribution of real human speech.

In conclusion, the main contribution of this paper is the successful synthesis of advanced AI research with user-centric application design. We have demonstrated that it is possible to not only correct the deep-level contextual errors in neural synthesis models but also to package this complex technology into a simple, reliable, and accessible tool. The resulting mobile application empowers users by breaking down language barriers, offering a robust method for clear and secure communication across multiple languages. This project serves as a complete validation of the artifact correction methodology, proving its value in a practical system that directly enhances human interaction.

References

1. S. Jeong, J. S. Sung, I. Hwang, and J. Choi, "An Automated Method to Correct Artifacts in Neural Text-to-Speech Models," IEEE Transactions on Audio, Speech and Language Processing, vol. 33, pp. 1375-1388, 2025.
2. J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, et al., "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), 2018, pp. 4779-4783.
3. A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," in Proc. SSW, 2016, p. 125.
4. Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.
5. J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high-fidelity speech synthesis," in Proc. NeurIPS, 2020, vol. 33, pp. 17022-17033.
6. J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in Proc. Int. Conf. Mach. Learn. (ICML), 2021, pp. 5530-5540.
7. C.-C. Lo, C.-Y. Hsieh, Y. S. W. H. S. Lee, and H.-Y. Lee, "MOSNet: Deep learning based objective assessment for voice conversion," in Proc. Interspeech, 2019, pp. 1541-1545.
8. A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), 2001, vol. 2, pp. 749-752.
9. T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila, "Improved precision and recall metric for assessing generative models," in Proc. NeurIPS, 2019, vol. 32.
10. A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Proc. NeurIPS, 2020, vol. 33, pp. 12449-12460.
11. Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, et al., "Tacotron: Towards end-to-end speech synthesis," in Proc. Interspeech, 2017, pp. 4006-4010.

12. N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in Proc. AAAI Conf. Artif. Intell., 2019, vol. 33, pp. 6706-6713.
13. X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," arXiv preprint arXiv:2106.15561, 2021.
14. P. K. Kakade and H. A. Murthy, "Syllable-based grapheme-to-phoneme conversion for text-to-speech synthesis in Hindi," in Proc. Interspeech, 2004.
15. S. S. R. V. B. Y. Ramana, "A common attribute based grapheme to phoneme converter for Indian languages," Int. J. Comput. Appl., vol. 1, no. 14, pp. 104-108, 2010.
16. A. Kumar, V. K. T, R. K. V, and Y. V. S. S. S. Prasad, "IITM-TTS: A text-to-speech synthesis system for Tamil," in Proc. 6th ISCA Workshop Speech Synth., 2007, pp. 266-271.
17. R. Jia, R. A. R. S, A. M. B, and S. K, "A unified text-to-speech framework for Indian languages," in Proc. Interspeech, 2018, pp. 3177-3181.
18. J. Postel, "Simple Mail Transfer Protocol," RFC 821, Aug. 1982.
19. E. L. R. T. K. A. G. S. and T. L. G. R. K, "A review of on-device neural network inference for mobile and embedded applications," IEEE Access, vol. 9, pp. 79659-79684, 2021.
20. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., "TensorFlow: A system for large-scale machine learning," in Proc. 12th USENIX Symp. Oper. Syst. Des. Implement. (OSDI), 2016, pp. 265-283.
21. Y. Yasuda, X. Wang, and J. Yamagishi, "Investigation of learning abilities on linguistic features in sequence-to-sequence text-to-speech synthesis," Comput. Speech Lang., vol. 67, p. 101183, 2021.
22. D. K. W. J. and A. K, "A study on the use of G2P converters for multilingual TTS in Indian languages," in Proc. O-COCOSDA/CASLRE, 2016, pp. 1-6.
23. J. Fong, D. Lyth, G. E. Henter, H. Tang, and S. King, "Speech audio corrector: Using speech from non-target speakers for one-off correction of mispronunciations in grapheme-input text-to-speech," in Proc. Interspeech, 2022, pp. 1213-1217.
24. Z. Borsos, M. Sharifi, and M. Tagliasacchi, "Speechpainter: Text-conditioned speech inpainting," arXiv preprint arXiv:2202.07273, 2022.
25. K. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Schuster, Y. Wu, and Y. Cao, "High-fidelity speech synthesis with adversarially trained conditional normalising flows," arXiv preprint arXiv:2006.07519, 2020.
26. W. Ping, K. Peng, A. Gibiansky, and S. O. Arik, "Deep Voice 3: Scaling text-to-speech with convolutional sequence learning," in Proc. Int. Conf. Learn. Represent. (ICLR), 2018.
27. P. C. M. L. S. T. R. and H. M. P. V, "A systematic review of security and privacy in mobile applications," IEEE Commun. Surv. Tutor., vol. 18, no. 1, pp. 76-101, 2015.
28. S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman, et al., "Deep voice: Real-time neural text-to-speech," in Proc. Int. Conf. Mach. Learn. (ICML), 2017, pp. 195-204.
29. S. K. K. K. P. J. K. and J. B, "A multi-speaker multi-style voice cloning," in Proc. Interspeech, 2021, pp. 4019-4023.
30. Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, R. Skerry-Ryan, and Y. Cao, "Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning," in Proc. Interspeech, 2019, pp. 2060-2064.